To appear in *Advanced Robotics*
Vol. 00, No. 00, January 2013, 1–11

A Cloud Robotics Approach towards Dialogue-Oriented Robot Speech

Komei Sugiura[a*] Yoshinori Shiga[a], Hisashi Kawai[b], Teruhisa Misu[a] and Chiori Hori[a]

[a] *National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika, Soraku, Kyoto 619-0289, Japan*; [b] *KDDI R&D Laboratories, 2-1-15 Ohara, Fujimino, Saitama 356-8502, Japan.*

(*January 2014*)

Robot utterances generally sound monotonous, unnatural, and unfriendly because their Text-to-Speech (TTS) systems are not optimized for communication but for text-reading. Here we present a non-monologue speech synthesis for robots. The key novelty lies in speech synthesis based on Hidden Markov models (HMMs) using a non-monologue corpus: we collected a speech corpus in a non-monologue style in which two professional voice talents read scripted dialogues, and HMMs were then trained with the corpus and used for speech synthesis. We conducted experiments in which the proposed method was evaluated by 24 subjects in three scenarios: text-reading, dialogue, and domestic service robot (DSR) scenarios. In the DSR scenario, we used a physical robot and compared our proposed method with a baseline method using the standard Mean Opinion Score (MOS) criterion. Our experimental results showed that our proposed method's performance was (1) at the same level as the baseline method in the text-reading scenario and (2) exceeded it in the DSR scenario. We deployed our proposed system as a cloud-based speech synthesis service so that it can be used without any cost. [1]

Keywords: human-robot interaction, speech synthesis, service robots, social robots

1. Introduction

Natural communication with humans is one of the most difficult challenges in human-robot interaction (HRI) studies. It requires sophisticated verbal and non-verbal interaction capabilities, including speech recognition/synthesis, dialogue management, and motion recognition/generation. Both integrating these components and improving each fundamental technology are crucial.

In this paper, we focus on natural and friendly synthesized speech for robots. Our target domain is service robots that are capable of speech communication. Speech communication with them continues to gain interest from research communities, especially at conferences and academic competitions such as RoboCup@Home[2], which focuses on mobile manipulation and HRI. We demonstrated speech communication combined with imitation learning[3] and object learning[4] at previous RoboCup@Home competitions.

Although the quality of corpus-based synthesized speech has greatly improved for text-reading, it remains unsatisfactory when applied to robots. In most cases, the synthesized speech of robots sounds monotonous, unnatural, or unfriendly because their Text-to-Speech (TTS) systems are not optimized for communication but for text-reading. In most TTS systems, their corpora are collected in a monologue style so that the synthesized speech lacks the live aspect of conversations. Such monotonous speech is obviously not desirable for emotional and conversational expressions, such as apologies, requests, or acknowledgments. Moreover, from our experience, monotonous intonation often prevents novice users from realizing that the robot is asking a

*Corresponding author. Email: komei.sugiura@nict.go.jp

[1] An earlier version of this work will be presented in IEEE ICRA 2014 [1].

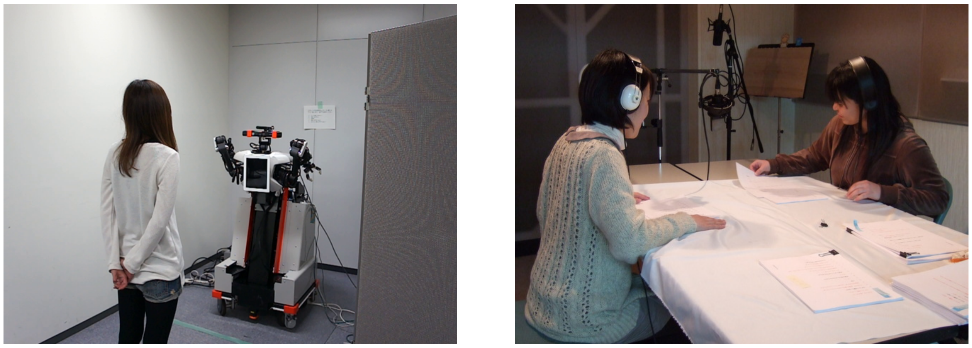

Figure 1. Left: Experimental environment used in DSR scenario. Right: Typical setting in non-monologue recordings.

question.

In this paper, we propose a cloud-based speech synthesis for service robots based on a non-monologue corpus. Unlike in other TTS systems, where a corpus is built in a monologue (text-reading) style, our method does not rely on monologue-style corpora. Insted we collected a non-monologue corpus, where two professional voice talents read scripted dialogues. We additionally collected a monologue corpus including phonetically balanced sentences. Separate HMMs were then trained with these corpora and used for speech synthesis. Our research problem is to what extent the non-monologue approach outperforms the conventional monologue approach.

The following are our key contributions:

- Monologue and non-monologue corpora were collected from professional voice talents and separately used for training two HMMs. The corpora are explained in detail in Section 3.
- Subject evaluations using the mean opinion score (MOS) criterion were conducted with a service robot (Figure 1-Left). In Section 5, we show the results that the non-monologue TTS outperformed the monologue TTS.
- We deployed our proposed system as a cloud-based speech synthesis service so that any roboticist can use it without any cost or authentication.

2. Related Work

Much research has attempted to improve the quality of emotional speech in robotics, conversation analysis, and speech synthesis studies (e.g. [5–7]). In the spoken dialogue systems (SDS) community, some recent studies have built expressive TTS systems for SDSs. In [8], HMMs were separately trained with speech data in different emotional tones such as liveliness, sulking, anxiousness, and relief. In the above studies, however, the recording was conducted in a monologue style. On the other hand, we propose a TTS system built from a dialogue corpus, which extends our previous work[9]. [10] also built a TTS system from a dialogue corpus, however they used a small corpus collected from non-professional speakers, so that the quality is not enough as a practical service. The main difference between the above studies and ours is that they did not conduct experiments with a physical robot, so that the contribution to the robotics is limited.

On the other hand, in robotics, [11] investigated the expression of emotion in synthesized speech for an anthropomorphic robot. However, most robotics studies did not conducted the standard task with the MOS (Mean Opinion Score) metric used in the speech synthesis community, which makes difficult to show contributions to the speech synthesis communities. Although our main contributions are to the robotics, we believe our method and system are very unique in the speech synthesis community.

In terms of practical robot applications, another issue exists. High-quality TTS systems are expensive and generally require more than several hundred mega bytes per voice font. However, storage and memory are limited in most robotic systems. A powerful solution is to use cloud resources (e.g., [12–14]). This approach is sometimes called “Cloud Robotics”[12] or “Cloud

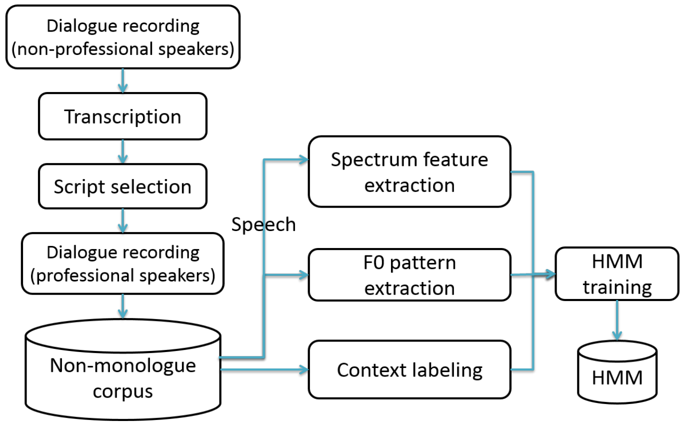

Figure 2. Schematic of our approach's procedure.

Networked Robotics"[13]. Although there exist commercial/non-commercial cloud services on speech synthesis, these are not optimized for communication with robots. This work is inspired by the above studies, however we focus on speech synthesis optimized for service robots.

3. Dialogue-Oriented Robot Speech

3.1 *Non-monologue Speech Synthesis*

In this paper, we propose a non-monologue speech synthesis for service robots. To build a non-monologue TTS system, we constructed a monologue corpus and trained HMMs with it. To obtain a high-quality TTS system based on HMM, data size and pronunciation consistency are of importance. For the data size issue, conventional studies on dialogue-style TTS had difficulty collecting large-scale, high-quality speech corpora. For example, the training data set sizes were 25 [min] in [10] and 558 sentences in [8]; however our maximum training data size was 433 [min] (14179 sentences). We will discuss the data size again in Section 6.

Another issue is the pronunciation consistency of the speakers. Even under identical phonetic context, the pronunciation of non-professional speakers is not consistent, which deteriorates the quality of the synthesized speech. For example, [10] built HMMs from a dialogue corpus of non-professional speakers contained in the CSJ corpus[15]. To the best of our knowledge, no TTS has been built from a corpus of face-to-face dialogues between professional voice talents whose pronunciation is highly consistent.

The schematic of the procedure in our approach is shown in Figure 2. First, we built a text corpus by transcribing the Kyoto Sightseeing Guidance Spoken Dialogue Corpus[16], which is a set of itinerary-planning dialogues in Japanese. It contains 160 hours of spontaneous speech data obtained from 328 pairs of tourists and professional guides. We extracted 21 balanced conversations to avoid one-sided examples and transcribed them as scripts in our recordings. Table 1 shows an example of the scripts. Although the original text was in Japanese, we show the translated text for readability and clarity.

Then we conducted non-monologue recordings (Figure 1-Right) from professional voice talents. In the recordings, they sat across a table and read the scripted dialogues naturally without overlapping each other because overlapped speech severely deteriorates the quality of synthesized speech. The recording was conducted in a soundproof room. The size of the non-monologue data was 466 [min] per person. The data were divided into the training and test data shown in Tables 3 and 4.

To compare the quality of our non-monologue approach with conventional monologue approaches, we also built a monologue corpus from the same voice talents who separately read scripts including phonetically balanced text in Japanese. The size of the monologue data per person was 180 [min], and 176-min-data of the total were used for training a monologue model.

We sampled the speech signals at 16 [kHz] with 16 bits/samples and computed 40-dimensional features. We used STRAIGHT[17] for extracting the spectral envelope and the F0. The feature vectors consist of 120 features including static, delta, and delta-delta parameters. Hidden semi-Markov models (HSMMs) were trained with these data using HTS[18]. The HSMMs had a 5-state left-to-right topology for modeling at the phone level. Phoneme segmentation was automatically conducted. For TTS we used the NX system, which replaces our previous system (XIMERA [19]).

3.2 *Cloud-Based Speech Synthesis*

We deployed our proposed method as a cloud-based system. The service is free for non-commercial use, and no authentication is required [1]. This service is language-independent so that users can write code in C++, Python, JavaScript, etc. and obtain synthetic speech by sending a JSON file (Figure 3) to the following URL: `http://rospeex.ucri.jgn-x.jp/nauth_json/jsServices/VoiceTraSS`. Then a .wav sound file encoded in base64 is returned to the user. The back-end system was developed for "VoiceTra," a speech-to-speech translation system[20]. Even though the service is multilingual, non-monologue speech synthesis is available only for Japanese.

```
{
   "method" : "speak",
   "params" : [
      "1.1",                            // command version
      {
         "language" : "ja",
         "text" : "INPUT_SENTENCE",     // text in utf-8
         "audioType" : "audio/x-wav",
         "voiceType" : "F128"           // voice-font ID
      }
   ]
}
```

Figure 3. Speech synthesis command format

We released a cloud-based service on Sept. 4, 2013. Table 6 summarizes the log information from Sept. 6 to 16. In the table, the "processing time" represents the time taken by the server to process requests. The real time factor, RTF, is defined as follows:

$$\mathrm{RTF} = \frac{T_p}{T_o}, \tag{1}$$

[1] A sample script for this service is available at `http://komeisugiura.jp/software/nm_tts.html`

Table 1. Example script in non-monologue corpus.

Guide	Hello.
Tourist	Hello.
Guide	My name is Yamamoto, I'll be helping you plan your trip to Kyoto.
Tourist	Okay.
Guide	I am pleased to help you today.
Tourist	Thank you.

where T_p and T_o denote the processing time and the length of the synthesized speech. From Table 6, we can see that the processing time is approximately 10% of the synthesized speech since the RTF ≈ 0.1.

Table 2. Summary of server logs

Number of unique IPs	813
Number of requests	7556
Processing time (median)	203 [msec]
Real time factor (median)	0.115

4. Experimental Setup

4.1 *General Setup*

We conducted forced-choice listening tests to compare the performance in terms of naturalness or friendliness between the baseline and proposed systems. The baseline system used a monologue corpus, and the proposed system used a non-monologue corpus. The synthetic sentences were evaluated by 24 subjects. Each age group (20-29, 30-39, or 40-49) consisted of four males and four females. All were native Japanese speakers. To avoid biased evaluations by specialists, we excluded from the subject groups researchers or students who are specializing in robotics or speech synthesis.

In the experiments, each subject first listened to an introductory speech synthesized by five systems: analysis-synthesis, baseline, and three variations of the proposed method. This process was required to acquaint them with the quality of both the synthesized and natural voices. Then they listened to the test-set sentences and rated their naturalness or friendliness on a scale of 1 to 5. The test-set sentences were not used for learning. The utterances were randomly presented within the experiment.

The systems under comparison are shown in Table 3. Although we recorded the speeches of two voice talents, we only used one of them throughout the subjective evaluation. For the baseline method, an HMM was trained with a monologue-style corpus by the same voice talent. Mono-176 and NonM-176 were prepared to compare our proposed and baseline systems in which they have the same amount of training sets. Non-M325 is bigger in size than NonM-176, however there is no major differences about the contents of these training sets. In terms of the contents of the training sets, the difference between NonM-433 and {Non-M176, Non-M325} is that NonM-433 contains "Tourist" utterances of the same voice talent.

We conducted experiments in three scenarios: text-reading, dialogue, and DSR. These scenarios were selected from the potential application fields, text-reading, dialogue systems, and human-robot interaction.

Table 3. Training sets

System	Recording style	Training set size
(0) AS (upper limit)	Analysis-synthesis	-
(1) Mono-176 (baseline)	Monologue	176 min. (2359 sentences)
(2) NonM-176 (proposed)	Non-monologue (Guide)	176 min. (4485 sentences)
(3) NonM-325 (proposed)	Non-monologue (Guide)	325 min. (8861 sentences)
(4) NonM-433 (proposed)	Non-monologue (Guide & Tourist)	433 min. (14179 sentences)

Table 4. Test sets

Scenario	Size of test-set	Contents
Text-reading	30 sentences	ATR503 J-set
Dialogue	12 dialogues	Tourist Guide
DSR	12 sentences	RoboCup@Home

4.2 *Text-Reading Scenario*

Table 4 shows the test sets. In the text-reading scenario, 30 sentences were synthesized by five systems. The sentences were selected from ATR503[21] J-set, which is a phonetically balanced corpus in Japanese. They were held out from the training set to evaluate the TTS quality for unknown text that is not contained in the training set.

We conducted a 5-point-scale MOS test on the naturalness of the speech:

5: very natural, 4: natural, 3: fair, 2: unnatural, 1: very unnatural.

The evaluation was conducted with headphones and a laptop PC. The subjects worked at their own pace.

4.3 *Dialogue Scenario*

In the dialogue scenario, we used 12 transcribed dialogues as the test set. We selected them from the Kyoto Sightseeing Guidance Spoken Dialogue Corpus[16] and they were not used for learning. In the dialogues, the "Guide" utterances were synthesized by each system, and the "Tourist" utterances were recorded by a voice talent.

The subjects listened to the dialogues and rated the naturalness of the guide utterances by the same 5-point-scale MOS. The same listening environment was used in this scenario as in the text-reading scenario.

4.4 *Domestic Service Robot Scenario*

In the DSR scenario, subjects listened to our robot platform called Daia (Figure 1-Left). The distance between the robot and the subject was 1 [m]. The ambient noise level at the subject's position was $L_{Aeq} = 66.7$ [dB], where L_{Aeq} represents the equivalent sound level. The main noise source was the robot itself, and the noise level was at almost same level compared to the RoboCup@Home situations. The SNR at the subject's position was 3.9 [dB].

Daia has a humanoid upper body (Kawada Industries HIRO), four omni-directional wheels (Neobotix Omni-Drive-Module), two laser range finders (Hokuyo UTM-30LX), a RGB-D camera (Microsoft Kinect), a directional microphone (Sanken CS-3e), a loudspeaker (Yamaha NX-U10), and two laptop PCs.

We conducted the experimental procedures in a Wizard-of-Oz style, where one experimenter controlled the utterance timing and another controlled the robot's gaze. The subjects rated the robot's friendliness on a 5-point-scale MOS test. In contrast to the other two scenarios, we did not evaluate the naturalness of the synthesized speech in this scenario. This was to avoid biased ratings based on the idea monotonous robot speech is natural as robot speech.

Table 5 shows the test-set sentences used in our experiment. The original text was in Japanese, but the translated text is shown for readability and clarity.

5. Results

5.1 *Similarity and Quality*

First, we define a similarity measure between the training and test sets. If the similarity is strong for the proposed method, the quality is obviously high. Then, we show the MOS results

Table 5. Test-set sentences used in the DSR scenario

I've learned the plastic bottle.
I'll go and get the object, is that okay?
I'll hold a plastic bottle.
I'll make some cotton candy now, tell me when it's ready.
I'll search another place for the object.
I'm moving to the living room, is that okay?
I'm sorry, I can't grab the object.
I'm sorry. I can't find the beverage.
The cotton candy is ready.
Time's up.
Would the next person please stand in front of me?
Yes, I understand.

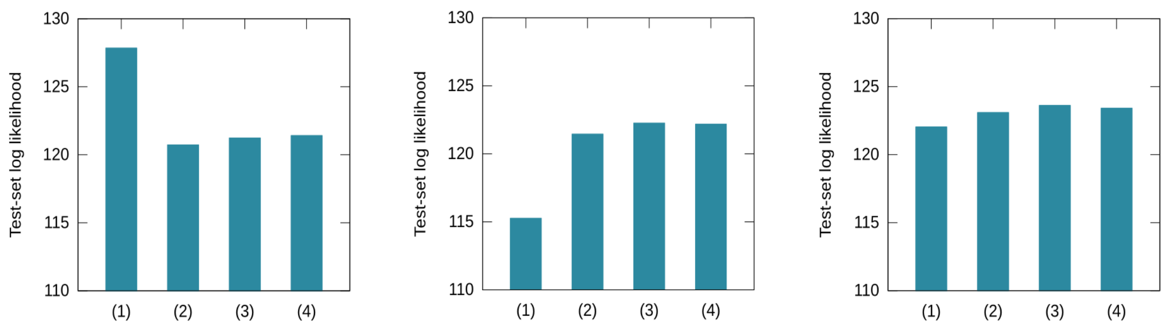

Figure 4. Comparison of test-set log likelihood: (1) Mono-176 (baseline), (2) NonM-176, (3) NonM-325, (4) and NonM-433. Left: text-reading scenario. Middle: dialog scenario. Right: DSR scenario.

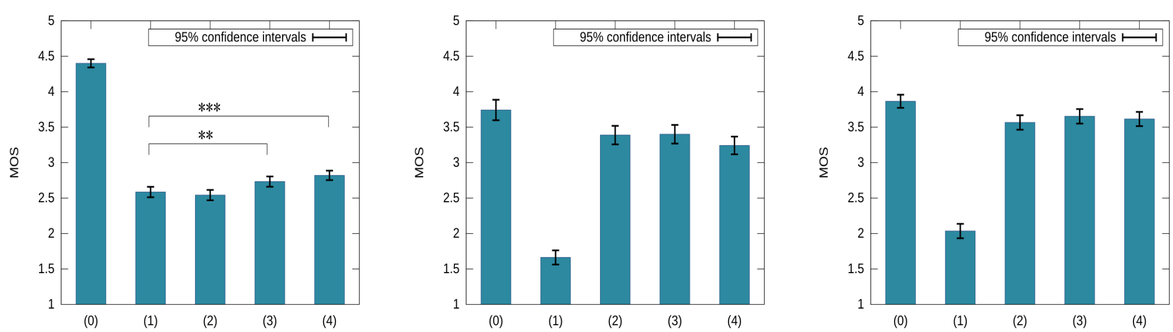

Figure 5. MOS results of speech quality: (0) Analysis-synthesis speech (upper limit), (1) Mono-176 (baseline), (2) NonM-176, (3) NonM-325, (4) and NonM-433. Analysis-synthesis speech shows the upper limit of the approach. If 95% confidence intervals do not overlap, the difference is significant. "**" and "***" represent $P < .01$ and $P < .001$, respectively. Left: text-reading scenario. Middle: dialog scenario. Right: DSR scenario.

to compare the quality of the synthesized speech. The proposed method outperforms the baseline even when the similarity is not strong.

Figure 4 compares the baseline and proposed methods in terms of the test-set likelihood. Test-set likelihood $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i}^{N} \frac{1}{T_i} \log p(\boldsymbol{O}_i | \boldsymbol{\lambda}_i), \quad (2)$$

where $N, T_i, \boldsymbol{O}_i$, and $\boldsymbol{\lambda}_i$ denote the number of samples in the test set, the number of frames of the ith sample, the observed speech of the ith sample, and the word sequence of the ith sample, respectively. Therefore, $\mathcal{L}$ can be regarded as the averaged log likelihood per frame. $\boldsymbol{O}_i$ was obtained from the read speech, and the word sequence was obtained from the speech's transcription.

5.2 *Text-Reading Scenario*

In Figure 5, the results of the opinion tests are shown. (0) represents the analysis-synthesis speech of the same voice talent, which was prepared to investigate the theoretical upper limit of our approach. The 95% confidence intervals visualize the statistical significance. If two of the intervals do not overlap, the difference is clearly significant. Additionally, statistical significance is shown by the "**" or "***" for cases where the difference is statistically significant but the intervals overlap.

The left panel of Figure 4 shows the similarity results in the text-reading scenario. The vertical axis shows the log likelihood per frame. The $\mathcal{L}$ of the baseline method was higher than the proposed methods. This is reasonable if we consider the similarity between the training and test sets. In this text-reading scenario, both the training and test sets were monologue-style corpora. Generally speaking, we can obtain higher test-set likelihood if the training and test sets are similar.

From the above result, it is reasonable to predict that the quality of the baseline method will exceed the proposed method in the MOS evaluation. However, in the left panel of Figure 5, no significant difference is shown between the baseline and proposed methods. This indicates that the proposed method's performance is at the same level as the baseline in the text-reading task. The figure also illustrates that NonM-325 and NonM-433 outperform the baseline method, showing that the non-monologue approach performs well for text-reading by increasing the amount of data.

5.3 *Dialogue Scenario*

In the middle panel of Figure 4, the similarity result of the dialogue scenario is shown. The proposed methods (NonM-176, NonM-325, and NonM-433) have higher likelihood values than the baseline method probably because the training sets of the proposed methods have strong similarities with the test set. Considering this fact, it is reasonable that the proposed methods clearly outperformed the baseline in the middle panel of Figure 5.

In the figure, NonM-433 has a lower MOS value than NonM176. Although this result is not statistically significant, the reader may be interested because better results are generally obtained when more data were used.

This fact can be explained as follows. Considering the practical situation where a service robot is talking to a human, we selected the "Guide" utterances as the test set. This is because the robots/systems are likely to give information to users in practical applications as a tour guide. In terms of the contents of the training sets, the difference between NonM-433 and {NonM-176, NonM-325} is that NonM-433 contains the "Tourist" utterances of the same voice talent. Therefore, NonM-433 did not obtain a higher MOS value than NonM-176 and NonM-325. This is also supported by the left panel of Figure 5 where NonM-433 outperforms NonM-176 and NonM-325. This is simply because NonM-433 contains more variations than NonM-176 which suffers from the over-fitting problem.

5.4 *Domestic Service Robot Scenario*

The right panel of Figure 4 shows the similarity result in the DSR scenario. Even though a difference is seen between Mono-176 and NonM-176, it is smaller than in the other scenarios.

The right panel of Figure 5 shows the MOS regarding friendliness in the DSR scenario. Proposed methods NonM-176, NonM-325, and NonM-433 have higher scores than the baseline method. This result clearly indicates that the proposed methods outperform the baseline method.

In the right panel of Figure 5, the proposed methods have large improvements from the baseline. What makes these improvements? A simple hypothesis is that the test-set utterances used in the DSR scenario have strong similarities with the training-set utterances of the non-monologue model. However, this does not explain the size of the difference in Figure 5; it is small in Figure 4. Another hypothesis is that the non-monologue aspect is the main reason for improving the quality of the synthesized speech in the DSR scenario.

Even though the results in this paper are promising, more analysis is required to clarify the mechanism behind the non-monologue approach. Future research includes analysis of the lexical and prosodic features that affect the quality as opposed to a monologue corpus. Despite the lack of such investigations, the results remain encouraging for research on natural interaction with robots.

6. Discussion

6.1 *Data Size*

In this subsection, we discuss the data sizes in terms of the number of subjects, training-set sentences, and tasks.

In this study, 30 synthetic sentences in the text-reading scenario were evaluated by 24 subjects. These numbers were selected based on the standard in speech synthesis studies. For example, 20 sentences were evaluated by six subjects in [10], and 25 sentences were evaluated by 16 subjects in [8].

Compared with other studies, the size of training set is sufficient, which is up to 443 [min] (14179 sentences). In contrast, most previous studies on dialogue-oriented TTS used very small data sets. Specifically, the sizes were 25 [min][10], 558 sentences[8], and 1200 sentences[22]. Generally speaking, it is difficult to build a dialogue-oriented TTS system with such a small data set.

In terms of task, these studies did not conduct experiments with a physical robot, so that the contribution to the robotics is limited. Some robotics studies (e.g. [11]) conducted HRI experiments with physical robots, but did not conducted the standard task with the MOS metric used in the speech synthesis community. On the other hand, this study compared the proposed and baseline methods in three scenarios with increasing difficulty in terms of interaction: text reading, dialogue and DSR. Even though the results in this paper are promising, more analysis is required to clarify the applicabilities to SDSs.

6.2 *Advantages/Disadvantages of Cloud-Based Approach*

From the viewpoint of speech synthesis, "Cloud Robotics"[13] has several advantages over stand-alone approaches.

- Collecting speech synthesis corpus for robots
  Currently there is no standardized corpus for robot dialogues, which makes difficult to compare methods.
- Intellectual properties
  Service providers do not have to distribute their highly valuable acoustic/language models.
- Maintenance
  Service providers do not have to ask users to apply for updates. Sometimes resolving dependencies is not easy in stand-alone systems.

The cloud-based approach has the following disadvantages:

- Demonstration under unstable networks

In many exhibitions (e.g., RoboCup), the network connections are not stable.

- Security
  The server can be attacked by malicious users.

We released a cloud-based service on Sept. 4, 2013. Table 6 summarizes the log information from Sept. 6 to 16. In the table, the "processing time" represents the time taken by the server to process requests. The real time factor, RTF, is defined as follows:

$$\mathrm{RTF} = \frac{T_p}{T_o}, \tag{3}$$

where T_p and T_o denote the processing time and the length of the synthesized speech. From Table 6, we can see that the processing time is approximately 10% of the synthesized speech since the RTF ≈ 0.1.

Table 6. Summary of server logs

Number of unique IPs	813
Number of requests	7556
Processing time (median)	203 [msec]
Real time factor (median)	0.115

7. Conclusion

In this paper, we presented a non-monologue speech synthesis for service robots. Conventional methods using monologue corpora have trouble synthesizing natural, conversational utterances. Our experimental results showed that our proposed method's performance almost approached the theoretical upper limit.

When building a conversational robot, it has been unavoidable to use a TTS system that cannot synthesize natural and friendly voices. Our experimental results, however, indicate that robot voices can be improved using non-monologue speech synthesis. One main contribution of this study is that we deployed our proposed system as a cloud-based speech synthesis service so that every roboticist can use it without any cost or authentication. Future direction will include quality improvement using collected logs. Demo video clips are also available: `http://komeisugiura.jp/video_gallery/`.

8. Acknowledgments

This work was partially supported by MEXT/JSPS KAKENHI Grant Number 24118710 / 24700188.

References

[1] Sugiura K, Shiga Y, Kawai H, Misu T, Hori C. Non-Monologue HMM-Based Speech Synthesis for Service Robots: A Cloud Robotics Approach. In: Proc. ICRA. 2014. p. 2237–2242.

[2] Iocchi L, van der Zant T. RoboCup@Home: Adaptive Benchmarking of Robot Bodies and Minds. In: Proceedings of the International Conference on Simulation, Modeling and Programming for Autonomous Robots. 2010. p. 171–182.

[3] Sugiura K, Iwahashi N, Kashioka H. Motion Generation by Reference-Point-Dependent Trajectory HMMs. In: Proc. IROS. 2011. p. 350–356.

[4] Nakamura T, Sugiura K, Nagai T, Iwahashi N, Toda T, Okada H, Omori T. Learning Novel Objects for Extended Mobile Manipulation. Journal of Intelligent & Robotic Systems. 2011;:1–18.

[5] Iida A, Campbell N, Higuchi F, Yasumura M. A Corpus-Based Speech Synthesis System with Emotion. Speech Communication. 2003;40(1):161–187.

[6] Tsuzuki R, Zen H, Tokuda K, Kitamura T, Bulut M, Narayanan S. Constructing Emotional Speech Synthesizers with Limited Speech Database. In: Proc. ICSLP. 2004. p. 1185–1188.

[7] Ishi CT, Ishiguro H, Hagita N. Analysis of acoustic-prosodic features related to paralinguistic information carried by interjections in dialogue speech. In: Proc. INTERSPEECH. 2011.

[8] Iwata K, Kobayashi T. Conversational Speech Synthesis System with Communication Situation Dependent HMMs. In: Proc. of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop. 2011. p. 113–123.

[9] Misu T, Mizukami E, Shiga Y, Kawamoto S, Kawai H, Nakamura S. Analysis on Effects of Text-to-Speech and Avatar Agent in Evoking Users' Spontaneous Listener's Reactions. In: Proc. of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop. 2011. p. 77–89.

[10] Koriyama T, Nose T, Kobayashi T. On the Use of Extended Context for HMM-Based Spontaneous Conversational Speech Synthesis. In: Proc. INTERSPEECH. 2011. p. 2657–2660.

[11] Breazeal C. Emotive Qualities in Lip-Synchronized Robot Speech. Advanced Robotics. 2003; 17(2):97–113.

[12] Kuffner J. Cloud-Enabled Robots. In: Proc. Humanoids. 2010.

[13] Kamei K, Nishio S, Hagita N, Sato M. Cloud Networked Robotics. Network, IEEE. 2012;26(3):28–34.

[14] Tenorth M, Perzylo AC, Lafrenz R, Beetz M. The RoboEarth Language: Representing and Exchanging Knowledge about Actions, Objects, and Environments. In: Proc. ICRA. 2012. p. 1284–1289.

[15] Maekawa K, Koiso H, Furui S, Isahara H. Spontaneous Speech Corpus of Japanese. In: Proc. of Second International Conference on Language Resources and Evaluation. Vol. 2. 2000. p. 947–952.

[16] Misu T, Ohtake K, Hori C, Kashioka H, Nakamura S. Annotating Communicative Function and Semantic Content in Dialogue Act for Construction of Consulting Dialogue Systems. In: Proc. Interspeech. 2009. p. 1843–1846.

[17] Kawahara H, Masuda-Katsuse I, de Cheveigné A. Restructuring Speech Representations Using a Pitch-Adaptive Time–Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. Speech Communication. 1999;27(3):187–207.

[18] Zen H, Nose T, Yamagishi J, Sako S, Masuko T, Black A, Tokuda K. The HMM-Based Speech Synthesis System (HTS) Version 2.0. In: Proc. of Sixth ISCA Workshop on Speech Synthesis. 2007. p. 294–299.

[19] Kawai H, Toda T, Ni J, Tsuzaki M, Tokuda K. XIMERA: A New TTS from ATR Based on Corpus-Based Technologies. In: Proceedings of Fifth ISCA Workshop on Speech Synthesis. 2004. p. 179–184.

[20] Matsuda S, Hu X, Shiga Y, Kashioka H, Hori C, Yasuda K, Okuma H, Uchiyama M, Sumita E, Kawai H, Nakamura S. Multilingual Speech-to-Speech Translation System "VoiceTra". In: Proc. Workshop on Field Speech and Mobile Data. 2013. p. 229–233.

[21] Kurematsu A, Takeda K, Sagisaka Y, Katagiri S, Kuwabara H, Shikano K. ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis. Speech Communication. 1990;9(4):357–363.

[22] Ling ZH, Richmond K, Yamagishi J, Wang RH. Integrating Articulatory Features into HMM-based Parametric Speech Synthesis. IEEE Trans Audio, Speech, and Language Processing. 2009;17(6):1171–1185.